SentiGAN: Generating Sentimental Texts via Mixture Adversarial Networks

Ke Wang & Xiaojun Wan
{wangke17,wanxiaojun}@pku.edu.cn

July 16, 2018

Institute of Computer Science and Technology, Peking University
Beijing , China

Outline

1. Introduction
2. Related Work
3. SentiGAN
4. Experiments
5. Conclusion and Future Work

Outline

Introduction

- Emotional intelligence is an important part of artificial intelligence.
  - Make machines more friendly to humans.
  - Make them look more intelligent.
- Challenges
  - Poor quality.
  - Lack of diversity.
  - Wrong sentiment.

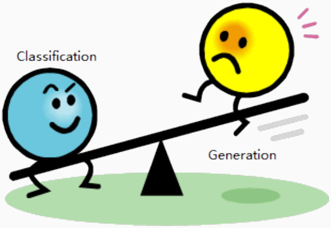

Figure 1: Sentiment classification is very strong, but the generation of sentimental texts does not.

Introduction

- Motivation
  - Use sentiment classifier to guide the generation of sentimental texts.
  - Use multi-class classification to make the text generated by the generator have more accurate sentiment.
- Contributions
  - We propose a novel framework SentiGAN to generate generic, diversified and high-quality sentimental texts of different sentiment labels.
  - We propose a new penalty based objective to make each generator in SentiGAN produce diversified texts of a specific sentiment label.
  - Extensive experiments are performed on four datasets and the results demonstrate the efficacy and superiority of our proposed model.

Outline

Related Work

- Unsupervised text generation.
  - Recurrent neural network language model.
    RNNLM
  - Generative Adversarial Nets and its variants.
    SeqGAN, RankGAN, LeakGAN, LabelGAN
  - Variational Autoencoders and its variants.
    VAE, semi-supervised VAE
- Others tasks
  Product review generation conditioned on specific inputs.

Outline

- Overall Framework
  1. Generator Learning.
  2. Discriminator Learning.

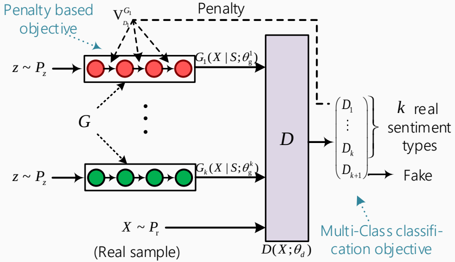

Figure 2: The framework of SentiGAN with k generators and one multi-class discriminator.

1. Generator Learning.
   We formalize the text generation problem as a sequential decision making process
   I Calculate the penalty:

$$V_{D_i}^{G_i}(S_{t-1}, X_t) = \begin{cases} \frac{1}{N}\sum_{n=1}^{N}\left(1 - D_i(X_{1:t}^n;\theta_d)\right) & t < |X| \\ 1 - D_i(X_{1:t};\theta_d) & t = |X| \end{cases} \tag{1}$$

   II Define the penalty based loss function:

$$L(X) = G_i(X_{t+1}|S_t;\theta_g^i) \cdot V_{D_i}^{G_i}(S_t, X_{t+1}) \tag{2}$$

   III Minimize the total penalty based value:

$$\begin{aligned} J_{G_i}(\theta_g^i) &= \mathbb{E}_{X\sim P_{g_i}}[L(X)] \\ &= \sum_{t=0}^{t=|X|-1} G_i(X_{t+1}|S_t;\theta_g^i) \cdot V_{D_i}^{G_i}(S_t, X_{t+1}) \end{aligned} \tag{3}$$

2. Discriminator Learning.
We use a multi-class classification objective that requires the discriminator to distinguish the real texts with each sentiment type and the generated texts.

$$\begin{aligned} J_D(\theta_d) = &- \mathbb{E}_{X \sim P_g} log D_{k+1}(X; \theta_d) \\ &- \sum_{i=1}^{k} \mathbb{E}_{X \sim P_{r_i}} log D_i(X; \theta_d) \end{aligned} \quad (4)$$

SentiGAN

- The adversarial training of generators and discriminator.
- We train them alternately.

Algorithm 1 The adversarial training process in SentiGAN

Input: Input noise, z; Generators, $\{G_i(X|S;\theta_g^i)\}_{i=1}^{i=k}$; Discriminator, $D(X;\theta_d)$; Real text dataset with k types of sentiment, $T = \{T_1, ..., T_k\}$;

Output: Well trained generators, $\{G_i(X|S;\theta_g^i)\}_{i=1}^{i=k}$;

Initialize $\{G_i\}_{i=1}^{i=k}$, D with random weights;
Pre-train $\{G_i\}_{i=1}^{i=k}$ using MLE on T;
Generate fake texts $F = \{F_i\}_{i=1}^{i=k}$ using $\{G_i\}_{i=1}^{i=k}$;
Pre-train $D(X;\theta_d)$ using $\{T_1, ..., T_k, F\}$;
repeat
    for g-steps **do**
        for i in $1 \sim k$ **do**
            Generate fake texts using $G_i(z;\theta_g^i)$;
            Calculate penalty $V_{D_i}^{G_i}$ by Eq (3) ;
            Update $G_i(z;\theta_g^i)$ by minimizing Eq (2);
        end for
    end for
    for d-steps **do**
        Generate fake texts $F = \{F_i\}_{i=1}^{i=k}$ using $\{G_i(X|S;\theta_g^i)\}_{i=1}^{i=k}$;
        Update $D(X;\theta_d)$ using $\{T_1, ..., T_k, F\}$ by minimizing Eq (5);
    end for
until SentiGAN converges
return ;

- The Multi-Class Classification Objective
  1. The optimal i-th generator can learn the distribution of the real texts with the i-th sentiment.

$$\begin{aligned}&\mathbb{E}_{X\sim P_g} log[\frac{P_g(X)}{P_{avg}(X)}] + \sum_{i=1}^{k} \mathbb{E}_{X\sim P_{r_i}} log[\frac{P_{r_i}(X)}{P_{avg}(X)}] \\ &-(k+1)log(k+1) \\ =&KL(\sum_{i=1}^{k} P_{g_i}(X)||P_{avg}(X)) + \sum_{i=1}^{k} KL(P_{r_i}(X)||P_{avg}(X)) \\ &-(k+1)log(k+1),\end{aligned} \tag{5}$$

  2. While keeping θ_d constant, the i-th generator aims to minimize the penalty ($V_{D_i}^{G_i}$) given by the discriminator.

- The Penalty-Based Objective
  1. Our penalty based objective can be considered as a measure of wasserstein distance which always provides meaningful gradients, even when the distributions of P_r and P_g do not overlap.

$$W(P_r, P_g) = \frac{1}{K} sup_{||L||_L \leq K} \mathbb{E}_{X \sim P_r}[L(X)] - \mathbb{E}_{X \sim P_g}[L(X)]. \tag{6}$$

  2. Our penalty-based loss function $G(X|S;\theta_g)V(X)$ can be thought of as adding $G(X|S;\theta_g)$ to the reward-based loss function $(-G(X|S;\theta_g)D(X;\theta_d))$.

$$\begin{aligned} G(X|S;\theta_g)V(X) =& G(X|S;\theta_g)(1 - D(X;\theta_d)) \\ =& G(X|S;\theta_g) - G(X|S;\theta_g)D(X;\theta_d) \end{aligned} \tag{7}$$

Outline

Experiments

- Simplify:
  We simply refer to the work of [Hu et al., 2017] and focus on generating short sentences (length $\leq$ 15 words) of two sentiment types (positive and negative).
- Datasets:
  - MR: Movie Reviews [Socher et al., 2013], contains 2133 positive sentences and 2370 negative sentences.
  - BR: Beer Reviews [Mcauley and Leskovec, 2013], contains 1437767 positive sentences and 11202 negative sentences.
  - CR: Customer Reviews [Hu and Liu, 2004], contains 1024 positive sentences and 501 negative sentences.

Experiments

- Baselines:
  - RNNLM[Mikolov et al., 2011]
  - SeqGAN[Yu et al., 2017]
  - Variational Autoencoders(VAE)[Kingma and Welling, 2014]
  - Conditional GAN(C-GAN)[Mirza and Osindero, 2014]
  - Semi-supervised VAE(S- VAE)[Kingma et al., 2014]

- Sentiment Accuracy of Generated Texts:

Accuracy	**MR**	**BR**	**CR**
Real Data	0.892	0.874	0.846
RNNLM	0.622	0.595	0.552
SeqGAN	0.717	0.684	0.632
VAE	0.751	0.721	0.643
SentiGAN(k=1)	0.803	0.750	0.731
C-GAN	0.822	0.773	0.762
S-VAE	0.831	0.793	0.727
SentiGAN(k=2)	**0.885**	**0.841**	**0.803**

Table 1: Comparison of sentiment accuracy of generated sentences. The real data is the training corpus.

- Quality of Generated Sentences
  - Fluency:
    We use a language modeling toolkit - SRILM [Stolcke, 2002] to test the fluency of generated sentences.

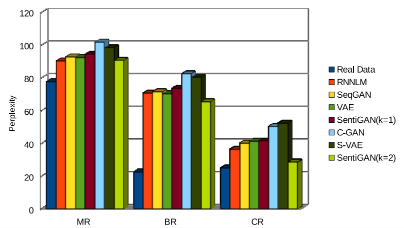

Figure 3: Comparison of fluency (Perplexity) of generated sentences (Lower perplexity means better fluency).

Experiments

- Quality of Generated Sentences:
  - Novelty:
    We want to investigate how different the generated sentences and the training corpus are.

$$Novelty(S_i) = 1 - \max\{\varphi(S_i, C_j)\}_{j=1}^{j=|C|}$$

Methods	**MR**	**BR**	**CR**
RNNLM	0.267	0.283	0.399
SeqGAN	0.298	0.328	0.437
VAE	0.287	0.347	0.417
SentiGAN(k=1)	0.344	0.409	0.479
C-GAN	0.368	0.398	0.482
S-VAE	0.328	0.369	0.437
SentiGAN(k=2)	**0.395**	**0.427**	**0.549**

Table 2: Comparison of the novelty of generated sentences.

Experiments

- Quality of Generated Sentences:
  - Diversity:
    We want to see if the generator can produce a variety of sentences.

$$Diversity(S_i) = 1 - \max\{\varphi(S_i, S_j)\}_{j=1}^{j=|S|, j \neq i}$$

Methods	MR	BR	CR
Real Data	0.753	0.705	0.741
RNNLM	0.691	0.677	0.663
SeqGAN	0.641	0.636	0.619
VAE	0.661	0.658	0.620
SentiGAN(k=1)	0.711	0.687	0.668
C-GAN	0.726	0.688	0.680
S-VAE	0.692	0.687	0.649
SentiGAN(k=2)	**0.741**	**0.713**	**0.708**

Table 3: Comparison of the diversity of generated sentences.

- Quality of Generated Sentences:
  - Intelligibility: We use human evaluation for evaluating the intelligibility of generated sentences.

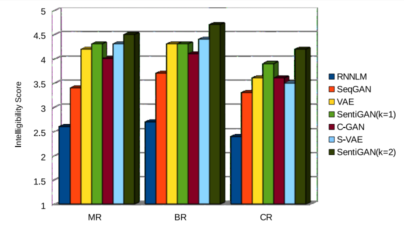

Figure 4: Comparison of intelligibility of generated sentences by human evaluation.

- Validation of Penalty-Based Objective

  We use a synthetic data set to test our proposed model in the mere use of the penalty based objective (i.e., SentiGAN(k=1)).

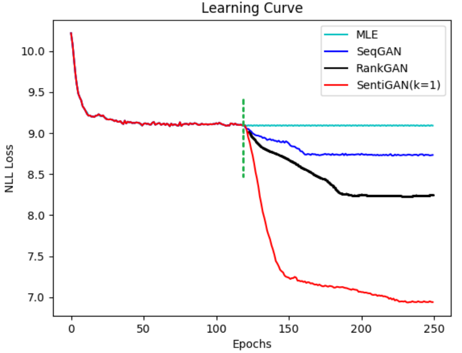

Figure 5: The illustration of learning curves. Dotted line is the end of pre-training.

- Case Study
  Our proposed model produces sentences that are more readable, sentimentally accurate, with better quality, and longer than that of C-GAN.

	SentiGAN(k=2)	C-GAN
Positive	a fantastic finally , simply perfect masterpiece.	give it credit , this is our 's brilliant . ***(Unreadable)***
	one of the greatest movies i have ever seen.	good , bloody fun movie
	funny and entertaining , just an emotionally idea but it was pretty good.	makes me smile every time to get on alien . ***(Unreadable)***
	the best comedy is a science fiction , captain is like a comic legend.	powerfully moving ! ***(Very short)***
Negative	one of the most disturbing and sickening movies i have ever seen.	very bad comedy. ***(Very short)***
	a story which fails to rise above its disgusting source material .	a mere shadow of its predecessors
	the comedy is nonexistent .	a timeless classic western dog ... ***(Wrong sentiment)***
	this is a truly bad movie .	one of those history movie traps

Figure 6: Examples sentences generated by SentiGAN and Conditional GAN trained on MR.

Outline

Conclusion and Future Work

- Conclusion
  - We propose a novel framework SentiGAN to generate generic, diversified and high-quality sentimental texts of different sentiment labels.
  - We propose a new penalty based objective to make each generator in SentiGAN produce diversified texts of a specific sentiment label.
  - Extensive experiments are performed on four datasets and the results demonstrate the efficacy and superiority of our proposed model.
- Future Work
  - Use of more complex and sophisticated generators.
  - Apply our model to generate texts with other kinds of labels (e.g., different writing styles).

Acknowledgments

- National Natural Science Foundation of China.
- Anonymous reviewers for their helpful comments.

Thank You!